Eine (sehr kurze) Einleitung in XML

Stelios Chronopoulos

Seminar für Griechische und Lateinische Philologie der Universität Freiburg

15. November 2018

2 Fragen zu Texten und 2 Methoden Antworten zu bekommen

- Frage 1: In einem Text will ich allen Personennamen erkennen, und sie als Liste extrahiere mit Hinweis auf die genaue Stelle im Text, in der jeder Name erscheint
- Frage 2: Ich will wissen, wie oft eine bestimmte argumentative Strutkur (z.B. p1 ist TRUE, p2 ist FALSE –> s ist TRUE) in einem Korpus von 200 Texten erscheint mit mit welchen anderen argumentativen Strukturen eventuell verbunden wird
- Zwei Methoden:
  - Inhalt/Text in einer relationellen Datenbank: ”strukturierte Daten”
  - Inhalt/Text als Datenbank: ”semi-strukturierten Daten”[fn:1]

Semi-strukturierten Daten

- ”Semi-structured data is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Therefore, it is also known as self-describing structure.”
- ”A data model (or datamodel) is an abstract model that organizes elements of data and standardizes how they relate to one another and to properties of the real world entities. For instance, a data model may specify that the data element representing a car be composed of a number of other elements which, in turn, represent the color and size of the car and define its owner.”

Semi-strukturierten Daten herstellen

- laufende Texte auszeichnen und Informationen kodieren
- die Strukturen eines Textes/Dokuments kenntlich machen –> der Computer kann sie erkennen und verarbeiten – ODER die Definition des Markups in der ”A Gentle Introduction to XML”: ”we define markup, or (synonymously) encoding, as any means of making explicit an interpretation of a text.”
- XML: eine Metasprache, die das Regelwerk leifert, um Texte auszuzeichnen –> eine Markup-Sprache

Hauptmerkmale von XML

- beliebig erweiterbar –> semantisch agnostisch
- deskriptives und nicht prozedurales oder präsentionales Markup
- universell: einfache Zeichendaten, Plattformunabhängig, offenes und voll-dokumentiertes W3C-Standard
- Maschine- und Menschenlesbar
- Dokument- und Datenzentriertes Markup

Trennung zwischen ”sein” und ”schein”

- Dokument d1 mit Markup getrennt von Dokkument d2 mit Informationen, die notwendig sind, um d1 maschinell zu verarbeiten (stylesheet)
- mehrere mögliche Art und Weisen d1 darzustellen / die in d1 beinhalteten Daten zu visualisieren (vgl. Realität in der Druckkultur)

Drei Fragen, die das Regelwerk beantworten muss

- wie unterscheidet sich Markup vom Text?
- welches Markup ist erlaubt?
- welches Markup ist undbeingt verlangt?

ein XML-Dokument muss / soll / kann:

- muss: syntaktisch korrekt sein: well-formed/wohlgeformte Dokumente
- soll/kann: valid sein: valid im Vergleich zu einem Schema

die Grundlage des Regelwerks

- eine, hierarchische Struktur
- Elemente
- Attribute

was genau wird ausgezeichnet?

- Sequenz aus Zeichen (und Leerzeichen), die
- strukturiert ist / Differenzierungen hat
  - auf der Ebene der sprachlichen Zeichen (analytische Einheiten)
  - auf der Ebene der graphischen Zeichen (physische Organisation des Textes auf dem tragenden Material)

Strukturen, Dokumenttypen und ihre Definitionen (DTD)

- Dokumente in Typen kategorisiert auf der Basis ihrer Strukturen
- Dokumenttypen haben Elemente in einer festen Anordnung
- deswegen können sie syntaktisch/strukturell überprüft und analysiert werden (parsing)

XML Elemente - gundlegende Einheit der Metasprache XML:

- abstrakte Darstellung von Textphänomemen
- Beschreibung von Textstrukturen
- Elemente und Elementennamen haben an sich keinen Sinn: ”XML legt nicht fest, welche Elemente und Elementnamen verwendet werden”

- Regel wie sie organisiert werden
  - jedes Element muss explizit markiert sein
  - jedes Element besteht aus einem <startTag> und einem </endTag>
  - zwischen diesen Tags steht der Inhalt des Elements (ein String)

Regel für wohlgeformte XML Dokumente

- eine Hierarchie: das root-element
- Elemente in Elementen aber keine partielle Überlappungen
- Anfang und Ende von Elementen muss immer explizit markiert sein

wirkliche Texte vs XML Regel

- eine XML Hierarchie – viele, verschiedene Textstrukturen: Konflikt?
- das Hauptproblem: verschiedene Ebenen in der Sprache und in der Textorganisation –> partiell überlappende Strukturen
- eine Hierarchie = eine strukturelle Ebene als primäre wählen zu müssen ist zu restriktiv - Notlösungen: <empty elements/>, pointing und linking

Attribute

- ”information that is in some sense descriptive of a specific element occurrence but not regarded as part of its content”
- Attattribute haben Werte und bilden ”atribute-value pairs”

Schemata

- XML bietet Flexibilität ABER wir wollen uns verständigen und austauschen können
- namespaces: ein Elementname gehört zu einer bestimmten, veröffentlichten Gruppe von Namen
- valides XML: ein Element e wird immer als <e></e> ausgezeichnet - das Schema: Namen von Elementen + bestimmte Regel (patterns) wie bestimmte Elemente kombiniert werden können/dürfen
- das Schema ist das Produkt einer Interpretation